Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

Lecture 6: ANOVA

Slides adapted from Cecilia Earls

Suppose under H0:

$$\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

t value

t-distribution degree of freedom = sample size -1

Critical value: t_α

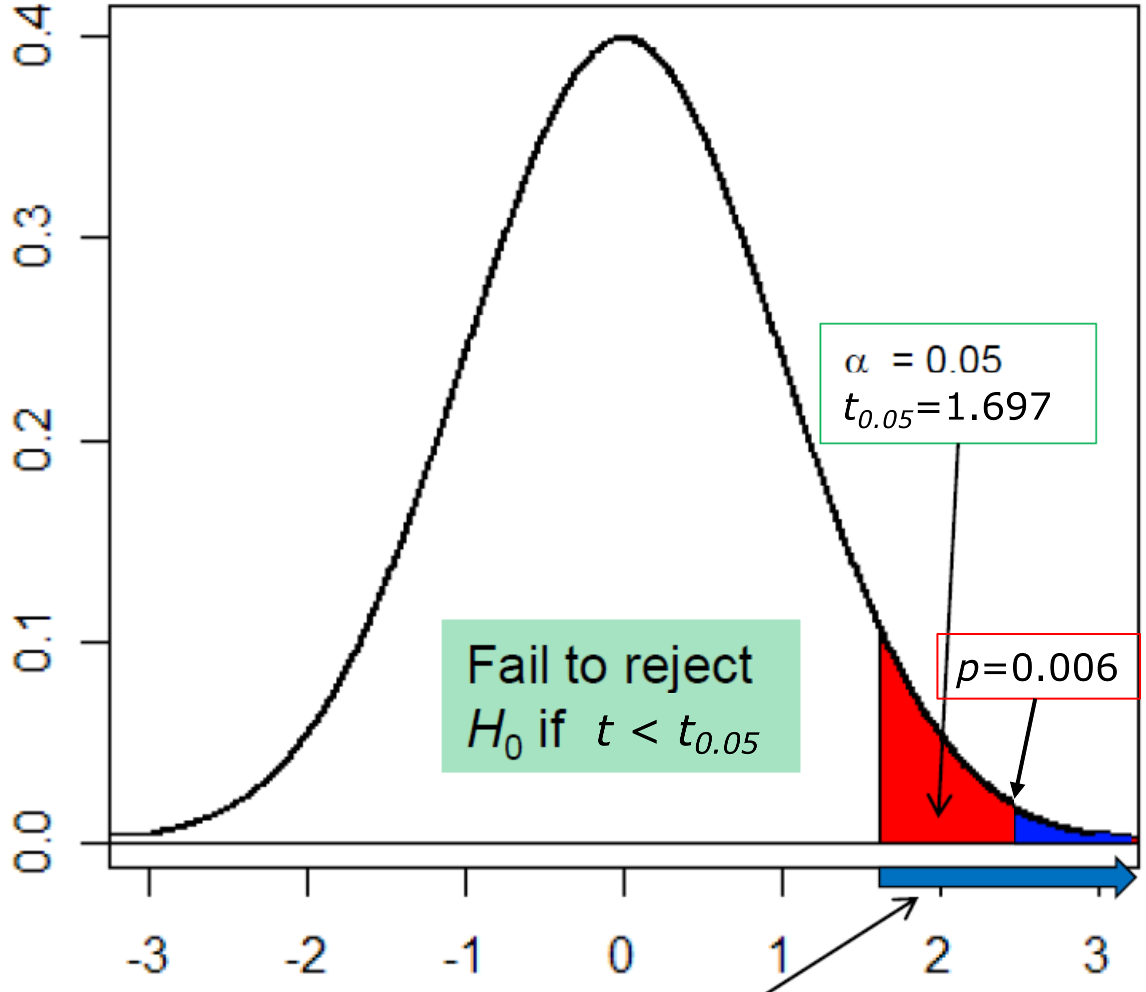
$(1-\alpha)$100th percentile of t_{n-1}

In R: $t_{0.05}$= qt(1-0.05,n-1)

p-value:

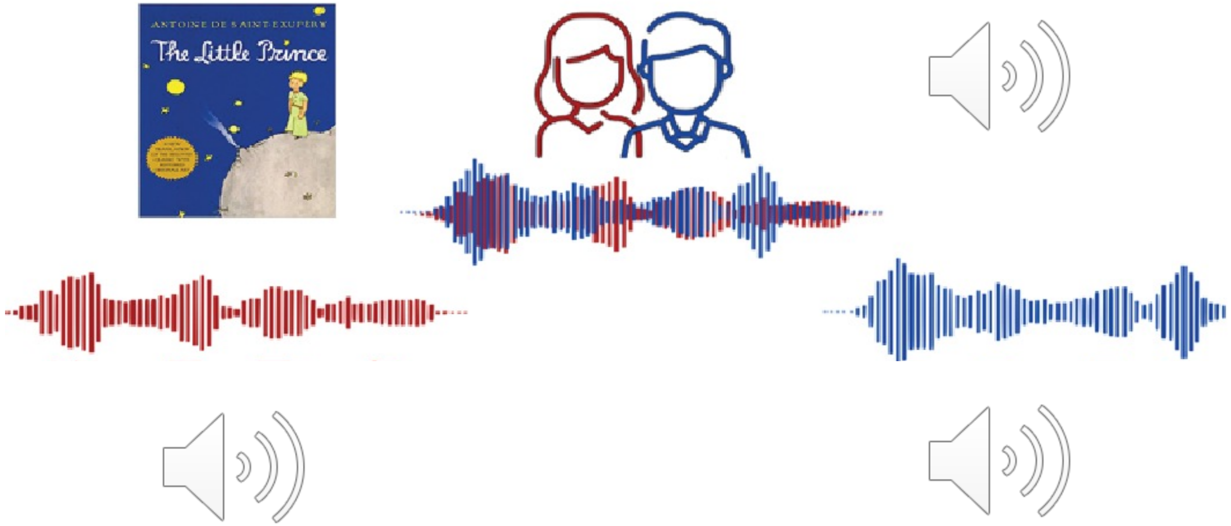
In R: *p* = 1-pt(*t*,n-1)

α = 0.05, one-sided



Reject H0 if $t > t_{0.05}$ or $p < \alpha$

Example: Cocktail party experiment



After listening, rate how clear the speech was from 1-5

1: not clear at all

5: very clear

Participants:

hearing-impaired adults (N=45): High-frequency hearing loss (> 8kz)

normal hearing adults (N=49)

Question 1: Is mixed speech more difficult to understand than single-talker speech for hearing-impaired and normal hearing listeners?

Question 2: Are hearing-impaired listeners have more difficulty understanding mixed and single-talker speech than normal hearing listeners?

Q1: within group

group	mixed	single
hearing-impaired N=45	M=3.44 SD=1.03	M=4.16 SD=0.85
normal N=49	M=3.94 SD=0.99	M=4.63 SD=0.64

Q2: between group

$$t^* = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

H_o: $\mu_1 - \mu_2 \leq D_0$ vs. H_a: $\mu_1 - \mu_2 > D_0$, reject H_o if $t^* > t_{\alpha, df}$

H_o: $\mu_1 - \mu_2 \geq D_0$ vs. H_a: $\mu_1 - \mu_2 < D_0$, reject H_o if $t^* < -t_{\alpha, df}$

H_o: $\mu_1 - \mu_2 = D_0$ vs. H_a: $\mu_1 - \mu_2 \neq D_0$, reject H_o if $|t^*| > t_{\alpha/2, df}$

$$\text{where: } t^* = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)} \text{ for } c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$$

More than two groups?

group	mixed	single
hearing-impaired N=45	M=3.44 SD=1.03	M=4.16 SD=0.85
normal N=49	M=3.94 SD=0.99	M=4.63 SD=0.64
children (Age:10-15) N=47	M=3.17 SD=0.92	M=4.06 SD=0.84

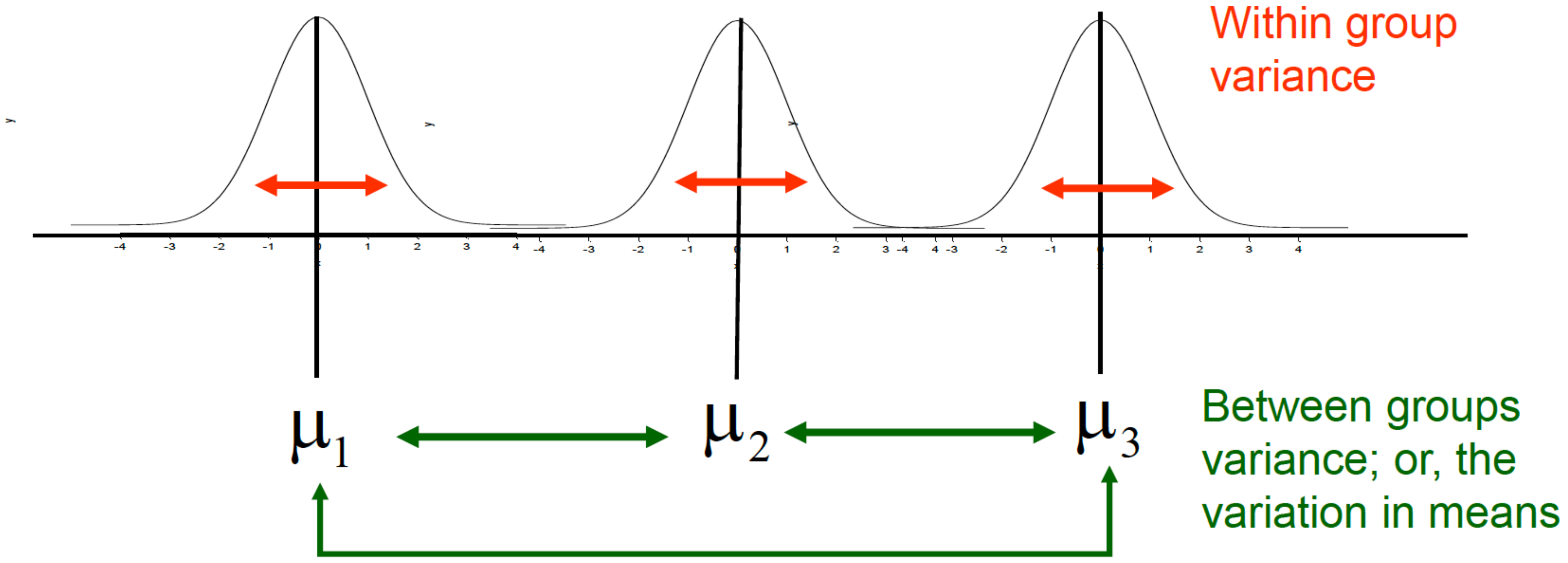
pair-wise *t*-tests:
children vs. normal vs. hearing-impaired, for both single and mixed speech
→ 6 *t*-tests

BUT:
Type I error rate: 0.05*6 = 0.3, **30% chance of false positives!**
Bonferroni correction: set $\alpha = 0.05 / 6 = 0.005$

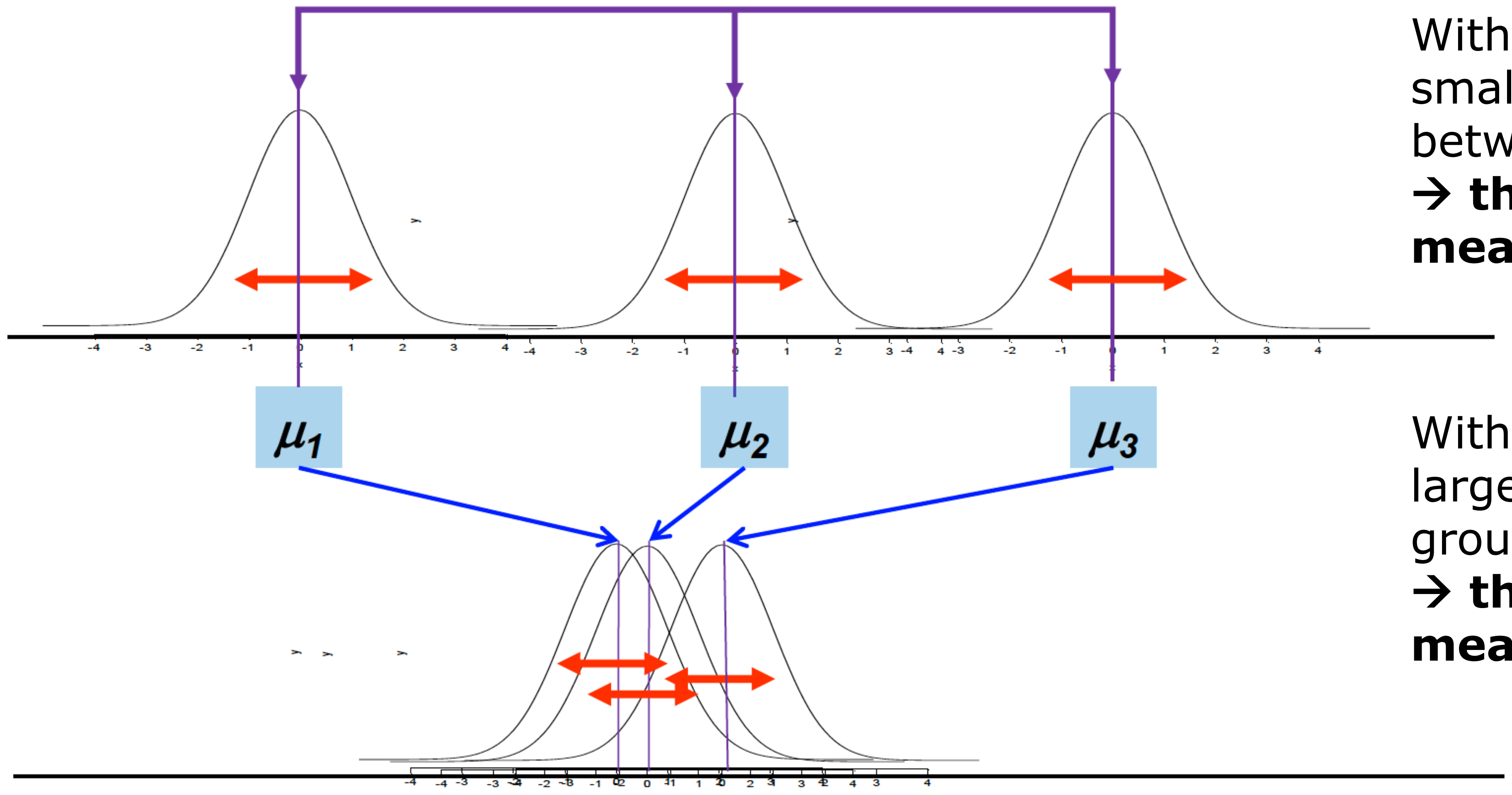
↓ **number of tests**

Better way: ANalysis Of VAriance (ANOVA)

Idea: ANOVA compares the variability between the group-specific means to a pooled estimate of "within group" variability, obtained using all observations in all groups.



One-way ANOVA



Within group variance is small compared to between groups variance.
→ **the group population means are different**

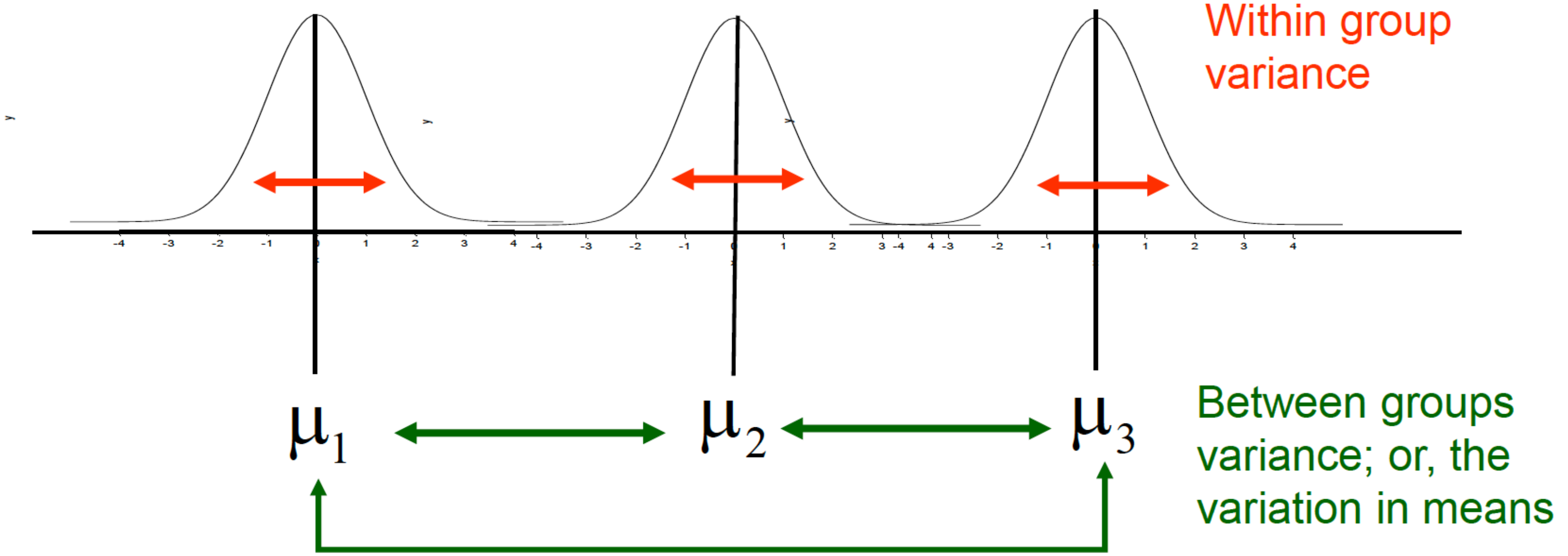
Within group variance is large compared to between groups variance
→ **the group population means are not different**

Perform a one-way ANOVA

1. **Compute sum of squares between group (SSB):**

$$\sum_{1}^{k} n_k(\overline{X_k} - \overline{X})^2$$

2. **Compute sum of squares within group / error (SSE):**
   sum of the squared differences between each individual observation and the group mean of that observation.
3. **Compute sum of squares total (SST): SSB+SSE**



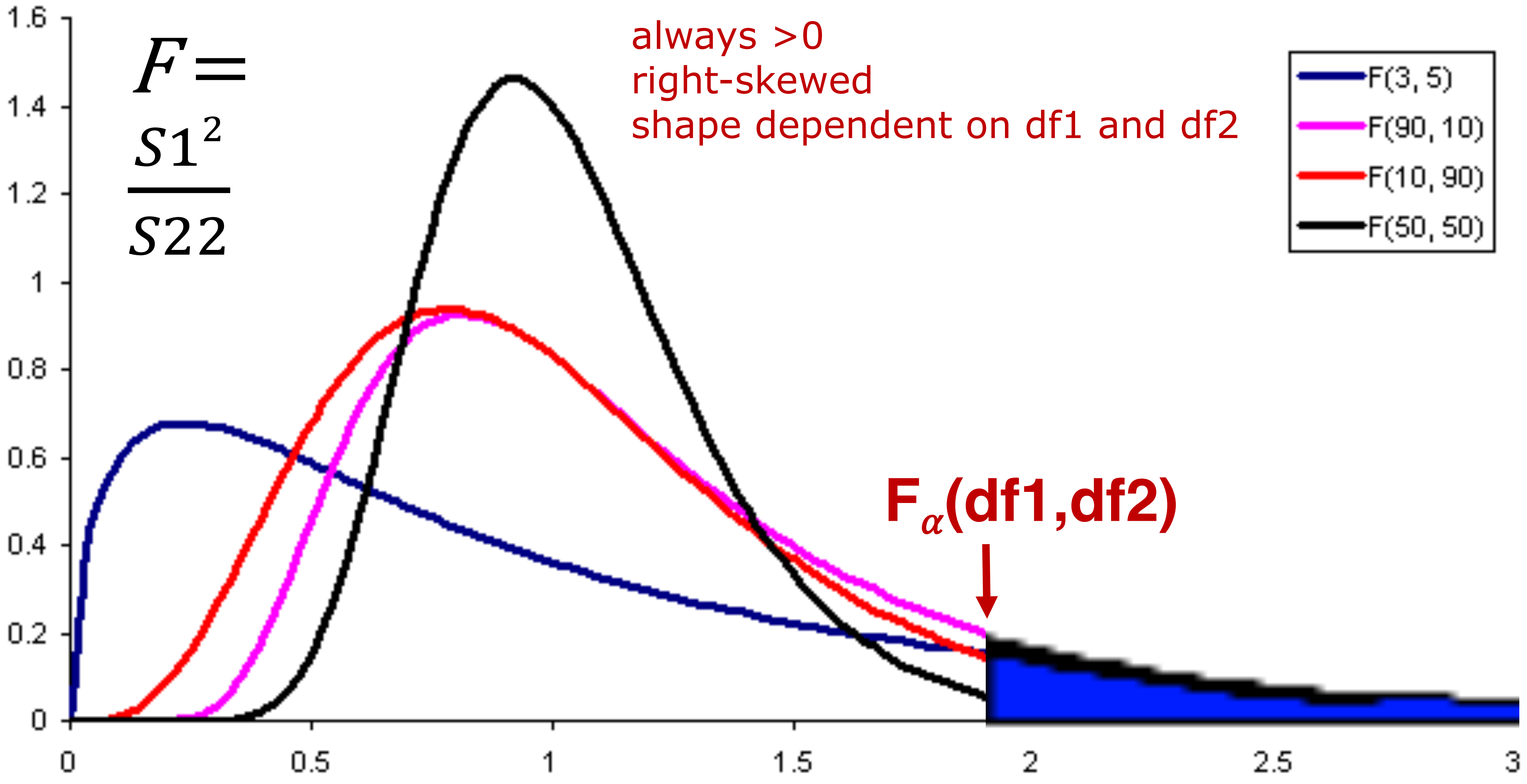
The ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean squares	*F* statistic
Between	k-1	SSB	MSB =SSB/(k-1)	MSB/MSE
Within (Error)	n-k	SSE	MSE =SSE/(n-k)	
Total	n-1	SST		

***k*:** number of groups
***n*:** total number of samples

Under H_0, F should tend to be close to 1. Under H_a, F should exceed 1, by an amount depending on both n and k.

F-distribution



Example: The cocktail party experiment

group	mixed	mean	grand mean
hearing-impaired	2,2,3	2.33	2.78
normal	2,4,4	3.33	
children	2,3,3	2.67	

SSB = $3*(2.33-2.78)^2 + 3*(3.33-2.78)^2 + 3*(2.67-2.78)^2 = 0.52$

SSE = $(2-2.33)^2 + (2-2.33)^2 + (3-2.33)^2$
$+ (2-3.33)^2 + (4-3.33)^2 + (4-3.33)^2$
$+ (2-2.67)^2 + (3-2.67)^2 + (3-2.67)^2 = 4$

k = 3, **n** = 9

MSB= SSB / (k-1) = 0.52/2 = 0.26

MSE = SSE / (n-k) = 0.67

F = MSB/MSE = 0.39

ANOVA in R

```
> cocktail_lm = lm(mix~group, data=cocktail)
> anova(cocktail_lm)
Analysis of Variance Table

Response: mix
           Df  Sum Sq Mean Sq F value    Pr(>F)
group       2  14.597  7.2987  7.5979 0.0007407 ***
Residuals 138 132.566  0.9606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```